AN EVALUATION OF THE EFFICIENCY OF A STANDARDIZED TEST OF LANGUAGE

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF EDUCATION

THE UNIVERSITY OF ALBERTA
FACULTY OF EDUCATION

BY
CECIL HENRY SANGSTER
MEDICINE HAT, ALBERTA

AUGUST, 1956

UNIVERSITY OF ALBERTA

FACULTY OF EDUCATION

The undersigned hereby certify that they have read and recommend to the School of Graduate Studies for acceptance, a thesis entitled, "An Evaluation of the Efficiency of a Standardized Test of Language" submitted by Cecil Henry Sangster, B.Ed., in partial fulfilment of the requirements for the degree of Master of Education.

August, 1956

SYNOPSIS

The primary purpose of this study was to determine the effectiveness of the California Language Tests, Elementary and Intermediate, as tests of English usage for selected Alberta pupils. Two randomly chosen samples were drawn from grades four and seven pupils of large urban and small urban centres in the province. The elementary test was administered to the grade four pupils and the intermediate to grade seven pupils.

The proportions of correct responses to each item of the tests were computed. Those recorded by the twenty-seven percent scoring highest and the twenty-seven percent scoring lowest, were employed in the item analysis. To determine the effectiveness of the test items, the techniques of item analysis devised by J. C. Flanagan and by Frederick B. Davis, were employed. Internal consistency indices for each item were tabulated. Distributions and patterns of item difficulty, in terms of percentage of successes, for both tests and their components, were illustrated graphically.

An examination of the data provided evidence to indicate that while some items of the California Language Tests, Elementary and Intermediate, exhibited relatively high discriminating power, in general, the internal consistency indices were sufficiently low to cast doubt upon the meaningfulness of the total score. The distributions of item

difficulty revealed that neither test is well adjusted to the levels of ability of the pupils sampled. The large proportion of relatively easy items in both tests, has the effect of reducing the power of the tests to differentiate between pupils at the upper levels of ability. Finally, the patterns of item difficulty were examined. In two subtests, considerations of item structure and responses compelled the authors to abandon any pretence at a pattern of ascending order of difficulty. In the remaining five tests, for our sample at least there is evidence of only nominal attention to the attainment of such a pattern.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER I

THE EVALUATION OF LANGUAGE ACHIEVEMENT

> Evaluation is that phase of education which is concerned with appraising the success of the school in attaining the purpose for which it is maintained by society.[1]

Evaluation is devoted to the measurement of changes in behavior brought about by a learning process. In a particular subject it should include attitudes, appreciations and work habits as well as knowledge and basic skills.

Scores from standardized tests, which are designed to measure progress toward predetermined objectives, enable the teacher to compare the progress of one pupil with that of another, or to compare the average score for a particular class with a norm. The fact that the average score for a class compares favorably with a particular norm does not mean only that the pupils are progressing as well as other similar pupils; it also means that they are doing just as much poor work and that they have as many unsolved problems.

Scales for the measurement of general achievement of written expression have proved to be reasonably satisfactory for the purposes of establishing standards. This information, while useful for administrative purposes, is of limited value

[1] John J. DeBoer, Walter V. Kaufers, and Helen R. Miller, Teaching Secondary English, New York, McGraw-Hill Book Company, 1951, p. 403

for the identification of pupil disabilities.

The fact that thirty pupils are seated in the same classroom and are assigned to the same grade is no guarantee of their homogeneity. Numerous studies have served to point out that each pupil is an individual, different and distinct in some way from others, and although many pupils successfully attain the objectives of a particular topic, each member of the class, at one time or another, experiences some type of learning difficulty, regardless of his ability. Such difficulties, which first make their appearance in the primary grades, frequently persist throughout the elementary grades, into the high school and beyond. Unless adequate diagnosis is made and appropriate remedial measures are carried out, many children are left with unidentified difficulties which continue to retard their progress in succeeding years.

Diagnosis is a logical process based on a consideration of all the available data concerning a particular individual. Fundamentally, diagnosis relates to the techniques by which one discovers the strengths and weaknesses of an individual as a basis for more effective guidance.

In the area of language a great deal of diagnosis may be carried out in the day-to-day observation of the pupil's written and oral expression. For adequate planning of regular language activities and remedial instruction, a systematic record of errors, in the form of a check list, is of great assistance. This should contain the objectives for

the grade and all preceding grades, simplified to meet the needs of the class. By means of such a guide, the teacher may check individual errors in composition at intervals throughout the term. Such an analysis frequently shows a few errors repeated over and over again. This discovery is of vital importance in planning the remedial program.

The task of recording errors might be lessened to great extent by the use of tests which provide a detailed diagnosis of important learning difficulties. Many diagnostic tests are accompanied by a diagnostic chart for the class and an individual chart or profile for the pupil. After recording the errors in a class chart, the teacher is in a position to determine which errors are general enough to require class treatment, and those which require individual or small group attention. The individual diagnostic profile might be kept for individual progress records or be given to the pupil to encourage self-appraisal.

Diagnosis in education, in its broadest sense, deals with a process that is characterized by change. The effects of each educational experience vary from pupil to pupil because of differences in maturity, experience and ability. The performance of an individual depends on the operation of many factors, such as the amount of effort he puts forth, his understanding of what is wanted, his environmental conditions and his interest in the activity. Variation in one or more of these factors results in a change in the

pupil's performance on a test if the same test is repeated. Nevertheless, if teachers use reliable objective techniques of diagnosis, they can in many cases determine the elements in the learning situation that should be corrected. In recent years there have been developed diagnostic tests which locate with a high degree of accuracy many of the faults that have previously been recognized in a vague way through observation of pupil responses. One can believe that it may be possible to develop tests of types of learning difficulty that will be as infallible as the well-known Schick test in medicine.

CHAPTER II

THE PROBLEM

In an analysis of the strengths and weakness of an individual or groups of individuals, the effectiveness of measurement depends primarily on the reliability and validity of the instruments use. When a test is to serve as a diagnostic instrument as well as a test of general achievement, it must yield component as well as total scores, each of which must be reliable and valid measures.

A test is reliable when it yields the same results consistently. If a group of persons taking the test, earn the same scores when the test is repeatedly administered, the test is measuring consistently and is said to be reliable. The consistency of the measurement is expressed in the reliability coefficient of the test.

While a reliable measurement is certainly desirable, the tester must also assure himself that the test used is valid. A test is valid when it actually measures what it purports to measure.

The validity of a test is best established by determining the correlation between the test scores and a criterion of known validity. When a criterion is not immediately available, an indirect method of correlating scores from the test in question with those of other established tests may be used for estimating the validity.

Test validity depends not only upon the validity of the content in general but also upon the validity of the individual items. The validity of the items, to a great extent, depends on the ability of the test author to select and to write objective items for each fact or idea to be tested. No amount of statistical work can make up for poorly written items or for a lack of editorial skill in preparing items for publication.

When items are selected for a test, a number of considerations must be taken into account. One of these is the level of difficulty of the items.

> The difficulty of an item may be defined as the proportion of a certain sample of testees that marks the item correctly, or it may be defined as the proportion of a certain sample of testees that actually knows the answer to an item.[4]

The solution to the problem of what constitutes the optimum level of item difficulty in an achievement test has not been agreed upon among test authors. Some authorities prefer approximately equal numbers of items at all levels of difficulty; others prefer to use a few easy and a few difficult items with the majority near the fifty per cent level of difficulty. There appears to be general agreement, however, that "the average difficulty of all items should be about 50 per cent".[5]

[4] Fredrick B. Davis, Item-Analysis Data; Cambridge Massachusetts, Graduate School of Education, Harvard University, 1946, p. 3

[5] H. E. Hawkes, E. F. Lindquist, and C. R. Mann, The Construction and use of Achievement Examinations; Boston, Massachusetts, Houghton Mifflin Company, 1936, p. 32

Items are not suitable for inclusion in a test if they are so easy that no pupil fails to respond correctly or if they are so difficult that no pupil is able to respond correctly.

The basic function of an achievement test is to place individuals along a scale of ability according to real differences in achievement. Such a function requires a test to discriminate between pupils of high and low levels of general achievement. The extent to which a test has this power is determined by the discriminating power of its items.

An item is said to have maximum discriminating power if every pupil who responds correctly ranks higher in general achievement than any student who fails on the item. An item has zero discriminating power when there is no difference in general achievement of the students who succeed on an item and those who fail. The ideal test would consist of items of high discriminating power distributed evenly over the difficulty scale.

Statistical techniques for expressing the discrimination index of an item require the use of an independent criterion measure of general achievement whose validity is very high. As it is usually difficult to secure a criterion of this quality, the total score on the test itself is frequently used as the criterion. When used in this way, the index really becomes an internal-consistency index. Such an index enables one to select items which are most effective in

measuring what the test as a whole measures. Items thus selected may be said to be valid only if the test itself is valid.

A low index may be the result of technical weaknesses in the composition of the test item. Ambiguities, irrelevant clues and "catch" questions make items ineffective. A low index of discrimination may also indicate insufficient knowledge on the part of the student. In the case of a multiple-choice question, a plausible but incorrect response may be the most frequent choice of the able and less able students alike as the result of their lack of understanding of the element being tested.

Finally, low indices may indicate that the "items are measuring rather specific points and are not highly correlated".[6] If the area being tested includes a type of achievement that is not highly related to other achievements within the same area, then items testing for that type of achievement would not be expected to yield high discrimination indices. These items, in spite of their low indices, should be retained if they are essential for adequate measurement of the area in question.

This immediately establishes a fundamental difference in the content of general achievement and diagnostic tests. A general achievement test is one designed to express in

[6] Davis, op. cit., p. 26

terms of a single score a pupil's achievement relative to others in the class. Its major purpose is to enable one to "rank" pupils in order of achievement and to interpret his score in terms of the norms provided for the test.

The meaningfulness of a single score as a measure of general achievement, depends upon the homogeneity of the field being tested. If low or negative correlations exist between specific types of achievement, then no single measure can adequately describe a pupil's achievement in all of these types simultaneously. Since there is no field of achievement, as usually defined in the present curriculum, in which all types of achievement are perfectly correlated, the use of a single general achievement test often will hide the fact that in certain types of achievement a pupil has deviated from his own general level.

The items which make up a general achievement test must be considered as representing a very restricted sampling of all items which might be selected on the basis of the subject involved. Items are restricted to the extent that only those which exibit reasonably high discrimination indices are included. It is only then that a single score has any meaning.

A diagnostic test, on the other hand, is one intended to discover specific deficiencies in learning within a particular field of achievement. Test items are selected for the purpose of sampling all areas of achievement within

the subject in question regardless of their correlation with one another. Their selection is based primarily upon logical rather than statistical considerations. Total scores, in such tests, have little meaning because of the heterogeneity of the area being tested and the lack of correlation between component scores. Part scores or the percentage of correct responses to homogeneous groups of items are the measures sought.

Test users have been encouraged to expect a single test to serve as a general achievement test and a diagnostic instrument at the same time. Many such tests have in fact been constructed. This may account for the failure to recognize the legitimate functions and essential characteristics of each type of test, as well as for the failure to realize that the two types cannot often be effectively combined in a single test.

The significance of a single score in a general achievement test depends upon the discriminating power of its items. It is desirable that the selection of items be restricted to those which correlate most highly with the total score so that a maximum in test homogeneity is achieved.

In a diagnostic test, items are selected to provide unitary sub-tests which measure independent areas of difficulty throughout the subject in question. Correlations between items and the total score, under such conditions, are frequently low or negative. Consequently, composite or total scores

have little or no meaning.

It is not possible to have it both ways. The correlation between items and total score, which makes for good measurement of general achievement, interferes with effective diagnosis. Likewise the lack of correlation between component scores in diagnostic tests weakens effective measurement of general achievement.

The rationale in achievement test construction is so completely at variance with that of diagnostic tests that adequate diagnosis from achievement tests is unlikely, due to the restricted sampling of items. Items are selected on the basis of their correlation with total score and are usually so few in number in essential diagnostic areas that component scores are unreliable. The use of diagnostic tests as general achievement measures is likewise unlikely because of the emphasis placed on unitary component tests. Items in these tests yield high correlations with their component scores but low correlations with a composite or total score. As a result, this heterogeneous measure produces a total score which is frequently both ambiguous and misleading.

CHAPTER III

STUDIES RELATED TO DIAGNOSTIC AND ACHIEVEMENT TESTING

1. Diagnostic Tests

Markedly different procedures must be employed in the construction of achievement and diagnostic tests due to their difference in purpose.

The principal purpose of achievement tests in education "is to enable us to rank pupils in a given group in the order of their total achievement within a given field."[5] Test results may be used in the classification of pupils for administrative purposes, the evaluation of teaching and the improvement of instructional methods. They also constitute an important preliminary to remedial teaching programs. In this connection, they are used in the identification of pupils with special disabilities and in measurement of progress in the course of remedial work.

The diagnostic test is intended, not to assess levels of achievement, but to reveal individual learning difficulties. It is constructed in such a way that each related skill within the subject measured in isolation, one at a time, until the nature of the pupil's weaknesses are revealed.

There are several types of diagnostic tests. Some tests, such as the Sangren-Woody Reading Test, seek to locate

[5] Hawkes, Lindquist and Mann, *op. cit.* p. 23

weaknesses by means of measurement of some of the specific phases of reading ability. Other tests, such as the Brueckner Diagnostic Tests in Fractions and Decimals, might be described as comprehensive inventories, the use of which enables one to locate a specific cause of difficulty for an individual.

In a sense almost any test may be called diagnostic. However, "many of the tests that are labeled diagnostic by their authors are in fact tests of general achievement."[6] We must determine the adequacy of the diagnostic cues provided by a test purporting to be diagnostic. It is easy to overstate the value of the diagnostic information provided by a particular test.

> To be truly diagnostic, a test must be based on a detailed analysis that permits the exact location of the spot in the work at which there is difficulty, or of the phase of general ability in which there is a deficiency.[7]

This requirement presents a serious problem to test authors. How is a diagnostic test, of reasonable length, to appraise component abilities with adequate reliability?

[6] Anne Anastasi, Psychological Testing, New York, The Macmillan Company, 1954, p. 22

[7] L. J. Brueckner and E. O. Melby, Diagnostic and Remedial Teaching, Cambridge, Mass., The Riverside Press, 1931, p. 74

> Tests generally have great difficulty in providing diagnoses which are reliable. The total score is often reliable, but decisions are made on the basis of sub-tests which are usually very short. Every sub-test is a test in itself and must demonstrate adequate reliability and validity. Moreover, since the diagnostic test is almost always intended to make important decisions about a particular individual, high reliability of sub-tests must be demanded.[8]

The reliability and validity of diagnosis depends upon the frequency of appearance of items measuring the same or related elements within the test. Errors detected on the basis of one or two items of a given type may or may not be due to carelessness or the chance selection of items involving specific elements of weakness. Foster E. Grossnickle, in a review of an arithmetic achievement test in the Fourth Mental Measurement Yearbook, states that "a diagnostic test must have at least three samples of a given type in order to provide a reliable diagnosis."[9]

Achievement examinations in the subject of language were among the first educational tests to be developed in America. In 1845, the Boston school committee was charged with the responsibility of making a yearly achievement inventory.[10] Previously this had been done by oral examinations

[8] L. J. Cronbach, Essentials of Psychological Testing New York, Harper and Brothers, 1949, p. 151

[9] Foster E. Grossnickle, "Review of the Arithmetic Essentials Test", The Fourth Mental Measurements Yearbook, Highland Park, N. J., Gryphon Press, 1953, p. 400

[10] O. W. Caldwell and S. A. Courtis, Then and Now in Education, 1845-1923, World Book Co., 1925, p. 6

but for the sake of the economy of time, written examinations were found necessary. Examinations were constructed for the subjects of arithmetic, astronomy, geography, grammar, history and natural philosophy.

In 1894, Dr. J. M. Rice[11] developed the first objective spelling tests in America. In 1912, a notable test, the Hilegas Composition Scale made its appearance.[12] This scale did much to stimulate interest in more accurate measurement of written composition.

> Despite the fact that language was one of the first school subjects in which experimental measurement in education was undertaken, progress in the development of measuring instruments in this field has not been particularly notable.[13]

This lack of progress in the development of adequate measuring instruments in language may be due, in part, to the complexity of the skills involved in written language and to the vagueness with which the objectives have been expressed.

[11] H. A. Greene, A. N. Jorgensen and J. R. Gerbich, Measurement and Evaluation in the Elementary School, Toronto, Longmans, Green and Co., 1947, p. 41

[12] Ibid. p. 365

[13] Harry A. Greene, "Measurement of General Merit", Encyclopedia of Educational Research, New York, The Mcmillan Co., 1952, p. 393

The California Language Test is among the most recently developed instruments for the measurement of language achievement. It is "the new edition of the diagnostic-survey instrument formerly called the Progressive Language Test".[14]

Test authorities have recognized many admirable qualities in the California Language Tests. Such features as the ease of administration and scoring, the achievement profile and the care taken in the preparation of the norms have been recognized by authorities as commendable. Other features of the tests have come under rather sharp criticism, most of which is directed toward the diagnostic power that the test is purported to possess.

In her book called "Psychological Testing",[15] Anne Anastasi describes the content and the organization of the California Achievement Tests including the California Language Test. She expresses doubt regarding the reliability of diagnostic categories containing as few as one or two items. Concerning the diagnostic features of the tests Dr. Anastasi states:

> Although of possible help in suggesting specific weaknesses in the individual's mastery of a skill, such an analysis may be quite misleading because of the small number of items involved. The reliabilities of the five major tests at each level appear to be adequate for

[14] Ernest W. Tiegs and Willis W. Clark, California Language Test Manual, Los Angeles, California Test Bureau, 1950, p. 2

[15] Anastasi, op. cit.

> survey purposes. No reliabilities are reported, however, for the sub-tests, some of which are very short. As for the further breakdown into functional elements, it is apparent that chance errors of measurement may play a considerable part in the results obtained in individual cases.[16]

R. L. Thorndike and E. Hagen, the authors of "Measurement and Evaluation in Psychology and Education",[17] comment on the pros and cons of fractionation in the California Achievement Tests. They admit that the use of sub-scores provides more specific information about what the individual is able to do. They point out, however, that fractionation is accompanied by shortening of the separate parts and that this shortening inevitably leads to lowered reliability in a diagnostic area where high reliability is most urgently needed. The result is that a test profile is created, with a number of deviations recorded, many of which are probably without significance.

Thorndike and Hagen suggest that a set of short and relatively diagnostic test scores might be used to advantage for a class rather than an individual. In the pooled results for a group, "chance errors of measurement tend to cancel out so that even a relatively unreliable test is adequate to bring out group differences."[18]

[16] Anastasi, op. cit., p. 469-470

[17] R. L. Thorndike and E. Hagen, Measurement and Evaluation in Psychology and Education, New York, John Wiley and Sons, Inc., 1955

[18] Ibid., p. 239

Reviews of the California Language Tests are written in "The Fourth Mental Measurements Yearbook"[19] by Robert C. Pooley and Gerald V. Lannholm.

Robert Pooley questions the validity and reliability of diagnoses based on the sub-tests and smaller diagnostic categories because of the few items involved. His main criticism is directed toward the limited content sampled by the items, particularly in the intermediate and advanced tests. He laments the fact that children are presumed to learn nothing more about language, other than grammatical facts, after the fourth grade.

Pooley suggests that an advance in the power to sustain an idea, improvement in sentence structure and an increase in the ability to organize and present materials logically are the skills on which pupils should be tested. He states that "since the California Language Tests ignore these basic factors, the use of the tests above the fourth grade can yield returns increasingly unrelated to the measurement of skill in language."[20]

Gerald V. Lannholm states that in his opinion "the tests will serve better as measures of general achievement than they will as diagnostic-analytic instruments."[21]

[19] O. K. Buros (Ed.), The Fourth Mental Measurements Yearbook, Highland Park, N. J., Gryphon Press, 1953

[20] Ibid., p. 149

[21] Ibid., p. 148

He bases this conclusion on his opinion that the California Language Tests do not adequately sample each of each of the elements of learning or skills involved.

The major criticisms of the California Language Tests appear to be directed toward two aspects of the tests, the reliability of the diagnostic categories and the content sampled by the tests.

There appears to be general agreement that diagnoses based on sub-scores and smaller diagnostic categories will be quite unreliable because of the few items measuring each skill. While this may be true for individual diagnosis, it is suggested that diagnostic scores will be of value for group diagnosis where chance errors tend to cancel out.

The authors of the California Language Tests have restricted the selection of items to those which measure "some of the most tangible and easily identifiable objectives of the curriculum."[22] Such skills as the use of capital letters and punctuation marks, and the recognition of complete sentences and parts of speech are measured in these tests. The ability to organize and present material logically, the ability to sustain an idea and an advance in sentence structure, are some of the major outcomes of effective language instruction which are ignored.

[22] Tiegs and Clark, op. cit., p. 4

2. Studies Related to Item Selection

Although considerable concern has been expressed by test authorities regarding the diagnostic value of the California Language Tests, little has been reported about their effectiveness as tests of general achievement. Some writers have remarked upon the inadequacy of the content but few have questioned the choice of items on the basis of their power to discriminate between levels of ability.

As the reliability and validity of a test depend upon the characteristics of the items, it follows that a test may be improved through the selection, substitution or revision of its items. With the growing refinement of test construction, the use of item analysis to improve test efficiency has received increasing attention.

A study conducted by John E. Anderson[23] illustrates the effects of item analysis upon the discriminative power of an English achievement test.

Following the administration and marking of an English examination, the tests were divided into three piles, an upper, middle and lower third on the basis of total score. The proportion of correct responses to each item for each third was computed and items which discriminated between

[23] John E. Anderson, "The Effect of Item Analysis upon the Discriminative Power of an Examination", Journal of Applied Psychology, Volume 19, 1935, pp. 237-244

groups were selected for subsequent examinations. Items of zero or negative discriminative power were removed from the test.

Of a total of 222 items, eighty-six of high discriminating power were selected as "good items", eighty-three of low discriminating power were selected as "poor items" and fifty-three items occupying intermediate positions were removed from the test. The pool of good items was referred to as a "good item test" and the poor items as a "poor item test". The tests were rescored to obtain the total score for each pupil. An analysis of the scores produced the following results.

1. The standard deviation and range of scores of the poor item test were considerably lower than those of good item test.
2. The scores from the good item test correlated .95 with the total test score.
3. The scores from the poor item test correlated .45 with the total test score.
4. The correlation between scores on the good and poor item tests was .21 .

Anderson states that "the fact that the reduced examination composed of good items, constituting approximately one-third of the total number of questions, correlates .95 with the total examination score, shows that the good items alone are almost as valuable as the total examination."[24]

[24] Ibid. p. 238

To arrive at an indication of the validity of the full test, the good item test and the poor item test, scores were correlated with final English grades, gradings on three book reports, scores on an Iowa English test and scores on the Minnesota College Aptitude Test.

Coefficients of reliability of the three tests were computed by the split-half method. The Spearman-Brown prophecy formula was then applied to determine the reliabilities of the good item and poor item tests if they contained 222 items.

Anderson concludes as follows:

1. The scores on the good item test predict the total test score with high accuracy and predict the final grade in English, book report grade and Iowa English scores better than does the total test.
2. Item analysis resulted in the production of a good item test which had almost as great reliability as the total test and which, if it had possessed the same number of items, would have had noticeably greater reliability.

A similar study was conducted by C. H. Lawshe Jr., and James S. Mayer[25] in which they compared two methods of item analysis to determine which produced the most reliable test. They also compared the reliability coefficients of

[25] C. H. Lawshe Jr., and James S. Mayer, "The Effect of Two Methods of Item Validation on Test Reliability", Journal of Applied Psychology, Volume 31, 1947, pp. 271-277

short tests composed of good items with that of a longer test composed of good and poor items.

An elementary psychology examination, containing 300 items, was used as the basis for the study. Two indices of discrimination were determined for each item by the correlation method devised by Flanagan and by the D-value method based on Lawshe's nomograph, adapted from Kelley's technique. The best 20, 40, 60, and 100 items were selected from the original 300 by the use of the two item analysis methods.

The reliabilities of the tests were computed by the split-half method and the following comparisons were made.

1. In selecting forty and sixty items, the two methods produced tests with no difference in reliability.
2. The Flanagan method produced the more reliable test of twenty items and the D-value method yielded the more reliable test of one hundred items.
3. The difference between reliability coefficients of the 100 item test produced by the Flanagan method and the original 300 item test, was not significant.
4. The difference between the reliability coefficients of the 100 item test, produced by the D-value method and the original 300 item test was significant at the 1 per cent level of confidence. The D-value test was the more reliable.

These studies indicate that through item analysis, it is possible to shorten a test and at the same time increase

its reliability. This is particularly true when items are selected on the basis of internal consistency.

Selecting items according to their internal consistency indices has the effect of conserving the items which have the highest intercorrelations. Consequently, the homogeneity of the test is raised and the ambiguity of the total score is reduced. In the measurement of general achievement, where the final score is of prime importance, the effectiveness of the test may be determined by the internal consistency indices of the items.

Reliability of measurement in achievement tests, as in diagnostic tests, depends largely upon the number of items in the tests or component tests and also upon the quality of the items. Component tests, as used in diagnostic instruments, should, like achievement tests, be homogeneous measures if their scores are to have meaning.

CHAPTER IV

RESEARCH DESIGN

1. The Purpose of the Study

This study was designed primarily to determine the effectiveness of the California Language Test items as general achievement test items for selected groups of Alberta pupils in grades four and seven.

This investigation is concerned with the evaluation of the test which was used to determine the levels of language achievement of various groups of Alberta pupils. The results of previous studies in this project have been reported by G. R. Conquest,[26] T. J. Reid[27] and Dr. H. T. Coutts and Dr. H. A. Baker. Reports of these studies may be found in the Alberta Journal of Educational Research, Volume 1, No. 2, June, 1955.

2. The Sample

For the larger survey of language achievement, a

[26] George R. Conquest, A Survey of English Language Achievement in Grades Four and Seven in Selected Alberta Schools, Unpublished Master's Thesis, University of Alberta, August, 1954.

[27] T. J. Reid, A Survey of the Language Achievement of Alberta School Children in Relation to Bilingualism, Sex and Intelligence, Unpublished Master's Thesis, University of Alberta, September, 1954.

[28] H. T. Coutts and H. S. Baker, "A Study of the Written Composition of a Representative Sample of Alberta Grade Four and Grade Seven Pupils", Alberta Journal of Educational Research, Volume 1, No. 2, June, 1955, pp. 5-18.

representative sample of 1,889 grade IV and VII pupils was selected from eight areas in the province. These areas were chosen on the basis of geographic location, racial origin of the population and socio-economic status. One town with a school population of between 250 and 1000 pupils was randomly selected from each area.

The graded and ungraded samples, selected from each of the eight areas, were approximately proportional to the total school population of these areas. The graded rural sample was restricted to towns with a total school population of less than 250.

The city of Edmonton was randomly selected for the large urban sample and Lethbridge for the small urban sample.

The present study has been confined to an anlysis of test results from a combined large urban and small urban sample. The grade IV sample contained 350 pupils (199 large urban and 151 small urban); while the grade VII contained 341 pupils (197 large urban and 144 small urban).

3. The Test Instrument

The tests used in this study were the California Language Test, Elementary, Form AA for the grade IV sample and the California Language Test, Intermediate, Form AA for the grade VII sample. These tests were developed jointly by Ernest W. Tiegs, Ph.D., and Willis W. Clark, Ed.D., and are published by the California Test Bureau - 5916 Hollywood Boulevard, Los Angeles 28, California.

On page 2 of the California Language Test manual the authors state that their test "is an instrument for accurately and objectively measuring pupil achievement in fundamental language skills. The test is standardized, and each item has been selected for its diagnostic value in measuring achievement in nineteen essential elements of language skill sampled in sub-test sections."[29] These tests are intended to measure only some of the basic skills of language and not the more complex factors which must be considered in assessing the quality of written expression.

The California Language Tests are revised editions of the Progressive Language Tests. The tests at the primary, elementary and intermediate levels are published in four forms, while the advanced level is published in three.

The California Language Tests are divided into two parts, Mechanics of English, and Grammar (Test Five) and Spelling (Test Six).

Test Five of the elementary test, intended for grades IV, V and VI, is divided into three sub-tests or sections: Capitalization - Section A, Punctuation - Section B, and Words and Sentences - Section C.

In the capitalization section, eight elements are tested, some of which are repeated so that the total possible

[29] Tiegs and Clark, op. cit., p. 2

score is 15. The first letter of three, four or five words in each line has a number above it. The number of the letter that should be a capital must be marked on the answer sheet. The extent to which a pupil over-capitalizes does not affect his score.

The punctuation section requires a period, comma, question mark or a quotation mark to be correctly placed in ten of a possible fifteen places in a story. There are five places where no punctuation is required. The score on this section is not affected by over-punctuation.

The section on words and sentences is divided into two parts. The first contains ten sentences in which a knowledge of number, case, tense and good usage is tested. The second part contains ten statements, some of which are complete sentences. The pupil is required to determine which statements are complete and which are not. Alternate response items are used throughout this section and no correction for guessing is made.

Test Six is a spelling test in which a pupil must select which, if any, of a set of four words is misspelled. There are thirty such sets in this test.

Test Five of the California Language Test, Intermediate, intended for grades VII, VIII and IX, is divided into four sub-tests or sections: Section A - Capitalization; Section B - Punctuation; Section C - Words and Sentences; and Section D - Parts of Speech.

The capitalization section tests the use of ten elements, two more than were tested in the elementary test. The total possible score is fifteen.

The intermediate punctuation test contains twenty items, five more than in the elementary punctuation test. The use of the period is not tested in the intermediate test. In its place the use of quotation marks within a quotation is tested.

The words and sentences section of the intermediate test has two items on singulars and plurals in addition to that which was tested in the elementary test.

Section D, Parts of Speech, is the only area in the intermediate test which is not similarly measured in the elementary test. This sub-test contains twenty multiple-choice items which require recognition of nouns, pronouns, verbs, adjectives, adverbs, conjunctions and prepositions.

The intermediate Test Six is a spelling test containing thirty multiple-choice items.

At the back of each test booklet, is an achievement profile on which a pupil's test results may be recorded and graphed. This profile may be used to illustrate a pupil's achievement in the mechanics of English, and grammar (as well as the component sub-tests), spelling and total language.

To the right of the achievement profile is a check-list called a Diagnostic Analysis of Learning Difficulties. This list is intended to be completed for those pupils who

fall below a desirable standard. It contains the elements presented in the test along with the item number which test that element.

The California Achievement Tests are considered to be power tests rather than speed tests. The authors state that the time limits of twenty-eight minutes for the intermediate test and twenty-five minutes for the elementary test are sufficient for about ninety percent of the pupils at each level. The remaining ten percent will have completed all the items they are capable of comprehending within these limits. In the event that a class completes a section before the specified time has elapsed, the examiner is instructed to proceed with the next section.

The coefficients of reliability and standard errors of measurement for the California Language Tests have been determined by the use of a Kuder-Richardson formula and are presented in the manual as follows.

TABLE I

RELIABILITIES AND STANDARD ERRORS OF MEASUREMENT FOR THE CALIFORNIA LANGUAGE TEST, ELEMENTARY AND INTERMEDIATE[30]

Test	Elementary (Gr. V)		Intermediate (Gr. VIII)	
	r	S.E. Meas.	r	S.E.
Mechanics of English, and Grammar	.90	.49	.89	.58
Spelling	.89	.45	.85	.71
Total Language	.95	.28	.93	.47

E. W. Tiegs and W. W. Clark, California Language Test Manual

[30] Tiegs and Clark, op. cit., p. 4

The authors of the California Language Tests state that their tests possess a high degree of validity, not only curricular but the more indirect cross-validation type as well. Coefficients of correlation between the Progressive Achievement Tests and seven other achievement test batteries are presented in a "Technical Report Supplementing Information Presented in the Manual of Directions".[31] The reader is reminded that the California Achievement Tests are revised editions of the Progressive Achievement Tests and that their validity may not be assumed on the basis of the validity of the Progressive Achievement Tests. Statistics supporting the validity of the California Achievement Tests are not reported in either the test manual or the technical supplement.

In the test manuals, the authors state that the items of the California Language Tests have been developed through four editions. The original selection of items followed a study of the objectives of most of the modern city and state courses of study. The items were then tried out in a number of areas throughout the United States, and those which were found to be of value were selected.

The authors further state that a number of studies have been made of the test items, the results of which have, with few exceptions, supported their choice. Where there

[31] Willis W. Clark, California Achievement Tests; A Technical Report Supplementing Information Presented in the Manuals of Directions, Los Angeles, California, California Test Bureau, 1951.

has been doubt as to the value of an item, it has been replaced by another.

Neither the test manual nor the technical supplement report statistically the results of the validation studies of either the test as a whole or of the individual test items.

4. Method of Investigation

In determining the effectiveness of the California Language Test items as general achievement test items, two related factors were investigated: (a) the discriminating power of the test items, and (b) the difficulty of the items.

At present, a wide variety of item analysis procedures is being used in test construction. The statistics most frequently employed as indices of item discrimination are the critical ratio, the product-moment coefficient of correlation, the phi coefficient of correlation, and the biserial coefficient of correlation.

The critical ratio provides an indication of how certain one may be that a particular item discriminates between high scoring and low scoring groups, but does not permit direct comparisons of amounts of discriminating power possessed by a number of items.

The phi coefficient is a product-moment coefficient computed from a fourfold table. Both the phi coefficient and the product-moment coefficient are unaffected by the size of the sample used for item-analysis purposes, but are

affected considerably by the proportion of the sample that answer the item correctly.

Items of fifty per cent difficulty tend to be characterized by the highest product-moment r's and phi coefficients. Consequently it is difficult to compare the discriminating power of items answered correctly by widely different proportions of the sample.

The biserial correlation method is considered by many as being the most satisfactory measure of item validity or internal consistency. It is the measure of the relationship "between each item and the total score, excluding the item in question".[32] This coefficient would probably find wider use if its computation was not so laborious.

To approximate the biserial correlation coefficient, with less labor, a method was devised by T. L. Kelly[33] in 1939. Kelly demonstrated that the lowest and highest twenty-seven per cent of a sample are the most serviceable groups for use in item-analysis, even though the items are not all of fifty per cent difficulty. He also outlined a method of estimating product-moment coefficients between items and the total test score using the highest and lowest twenty-seven per cent as criterion groups.

[32] Davis, _op_. _cit_., p. 9

[33] T. L. Kelly, "The Selection of Upper and Lower Groups for the Validation of Test Items", _Journal of Educational Psychology_, Volume XXX, January, 1939, pp. 17-24

In 1939, J. C. Flanagan, using the method described by Kelly, devised and discussed the use of a "Chart Showing the Values of the Product-Moment Coefficient of Correlation in a Normal Bivariate Population Corresponding to Given Proportions of Successes".[34] While the coefficients taken from Flanagan's chart are unaffected by item difficulty, they do not represent a linear function and therefore cannot be added, subtracted or averaged. This does not present a serious problem if one is interested only in comparing the discrimination power between items, but it does become a major consideration when the average index of discrimination for a number of items is required.

To satisfy the requirement for an index of discrimination which is unaffected by item difficulty and is a linear function, Fredrick B. Davis prepared an item analysis chart containing item discrimination indices derived from Flanagan's product-moment coefficients.

Flanagan's coefficients, which range from -.93 to +.93 were transformed into Fisher's z's by the use of the following formula:[35]

$$z = 1/2 \ \log_e \left[(1+r) - \log_e (1-r) \right]$$

[34] John C. Flanagan, "General Considerations in the Selection of Test Items and Short Method of Estimating the Product-Moment Coefficient from Data at the Tails of the Distribution", Journal of Educational Psychology, Volume XXX, 1939, p. 678.

[35] Davis, op. cit., p. 11

Davis then multiplied the z values, which represent a linear function, by a constant 60.241 (to five significant figures) to obtain item discrimination indices having a range from -100 to +100.

The Davis Item Analysis Chart provides both item discrimination indices and difficulty indices based on the proportion of successes in the highest and lowest twenty-seven per cent of the sample. The difficulty indices range from 1 to 99 and constitute a linear scale.

The Flanagan method and the Davis Method, which require essentially the same arrangement of data, were used in this study. The use of Flanagan's r's permitted the computation of a coefficient of forecasting efficiency for each item.

The writer was not primarily interested in exact differences of item difficulty. Instead, the general pattern of item difficulty and the average level of difficulty of sub-test and groups of items were the measures sought. Because of their ease of interpretation, the percentage of successes on each item was used instead of the Davis difficulty indices.

To arrive at these statistics it was necessary to make a count of the number of correct responses on each item for each member of the large and small urban sample at the grade four and grade seven levels. From this, the percentage of successes for each item was computed. Distributions of these data were compiled and histograms constructed for the

total test and the sub-tests. The pattern of item difficulty in terms of percentage of successes was graphically illustrated.

By the use of the forementioned data, the writer examined the elementary and intermediate tests for ineffective items and to determine how well the tests are adapted to the language abilities of samples of Alberta grade four and grade seven pupils.

CHAPTER V

ITEM-ANALYSIS DATA

1. Computations of the Proportion of Successes

The item-analysis techniques described by Davis and Flanagan require computation of the proportion of successes on items in the twenty-seven per cent of the sample scoring the highest on the criterion and the twenty-seven per cent scoring the lowest.

The proportion of successes is defined as the number of testees that know the answer to an item divided by the number of testees that respond to the item. The reader's attention is drawn to the fact that the word "knows" is used in the definition rather than the word "marks". This distinction is desirable whenever multiple-choice items are used because a certain number of testees, without knowing the answer, will mark the correct one by chance. To make allowance for this fact, a correction for chance has been employed by subtracting from the number of testees who mark an item correctly, the number of testees who mark the item incorrectly, divided by one less than the number of choices in the item. The proportion of successes thus becomes:[36]

$$P = \frac{R - \frac{W}{K-1}}{N_R}$$

[36] Davis, *op. cit.*, p. 6

where: $R =$ the number of testees that answer the item correctly.

$W =$ the number of testees that answer the item incorrectly.

$K =$ the number of choices in the item.

$N_R =$ the number of testees that respond to the item.

Following the scoring of the test papers, a tally was made of the responses to each item for the combined large urban and small urban samples on each of the elementary and intermediate tests. The responses were classified as correct, incorrect or no response. From these classifications the proportions of successes were computed.

In the computation of the proportion of successes for items in Sections A and B of both tests, no correction for chance was made. In Section A of the elementary test, the directions to the pupils state, "Mark the number of each letter that should be a capital. Some lines may have no such letter. If a line has no letter that should be capitalized, leave it blank." In Section B - Punctuation, the directions are essentially the same. Of four punctuation marks, the pupil must determine what marks, if any, are needed for each item. Consequently, on a four-choice item, the chances that the correct response is selected is 1:16.

The scoring procedure reduces the factor of chance still further by allowing credit for an item when a testee may select any number of the possible responses, provided that the correct response is included in the selection. It is possible, therefore, for a testee to receive a perfect score

by simply filling-in all responses spaces. No penalty is imposed for over-capitalization or over-punctuation.

Since the indiscriminate application of capital letters and punctuation marks indicates a lack of knowledge of their use, it is surprising to find that pupils are not penalized for over-capitalization and over-punctuation.

In Section B, a place is provided for the testee to indicate that no punctuation is necessary but such is not the case for Section A - Capitalization. Here, no capitalization is indicated by no response to an item. This results in an ambiguity in the interpretation of responses. No response may indicate that the testee refrained from marking an answer because of a lack of knowledge or a lack of time. Or it might be interpreted to mean that the testee, after reading the item, decided that no capitalization was necessary.

The test authors might have avoided this ambiguity by providing a place in Section A where the testee could indicate positively that no capitalization was needed.

For the purposes of this study, it was assumed that a <u>no</u> response on the last items of Section A was a result of insufficient time for the particular testee to complete the item. No response on an item, at the beginning or between other marked responses, was assumed to indicate that the testee decided that no capitalization was necessary. Such responses were included in N_R of the following formula.

$$P = \frac{R}{N_R}$$

Where: P = the proportion of successes.

R = the number of testees that respond to the item correctly.

N_R = the number of testees that respond to the item.

This formula results from subtracting the correction for chance from the formula quoted on page thirty-seven.

Since both the elementary and intermediate tests are constructed in the same manner, they share weaknesses in the method of scoring of Sections A and B.

2. Item Discrimination Indices

Since both the Davis and Flanagan methods of item analysis require a computation of the proportion of successes in the highest and lowest twenty-seven per cent of the sample, a single assembly of data was sufficient for both methods. Following the computation of the proportion of successes, item analysis charts were used to determine the discrimination indices.

The item-analysis chart, devised by Davis, is arranged in such a way that the percentage of successes of the highest twenty-seven per cent of the sample is located on the horizontal axis. The percentage of successes of the lowest twenty-seven per cent is located on the vertical axis. By reading horizontally and vertically from the previously determined percentages, two indices may be located simultaneously, one immediately above the other. The upper index

is the discrimination index and the lower, a difficulty index.

The Flanagan chart is constructed in a similar manner with the proportion of successes of the highest twenty-seven per cent placed on the horizontal axis, the proportion of successes of the lowest twenty-seven per cent is on the vertical axis. Product-moment coefficients of correlation may be located in the same manner as are the Davis indices of discrimination. The coefficients are given in multiples of five with each of the same value and sign being joined by a line. A coefficient, other than one of a multiple of five, must be estimated by its location between two adjacent lines. No difficulty indices are to be found in the Flanagan chart.

Coefficients of forcasting efficiency (E) which provide an estimate of the predictive efficiency of an obtained "r", were derived from Flanagan's product-moment coefficients as follows:[37]

$$E = 1 - \sqrt{1 - r^2}$$

where: E = the coefficient of forcasting efficiency

r = the product-moment coefficient of correlation

To illustrate the application of E, suppose that the correlation of an item with the total score is .45 . Then from the formula E = .11, and the item's efficiency in predicting the final score is said to be 11 per cent. This

[37] Henry E. Garrett, Statistics in Psychology and Education, New York, Longmans, Green and Company, 1947, p. 337

means that in 11 per cent of the cases, pupils responding correctly to an item whose correlation with the total score is .45, will receive higher total scores than those who respond incorrectly.

Table II contains item-analysis data based on the test results from the California Language Test, Elementary. The data in Table III are based on the results from the Intermediate test.

In both Table II and Table III, R refers to the number of pupils responding to the item, and P the proportion responding to the item correctly.

Items 5, 14, and 28 in Table II and 6, 12, 13, 20 and 24 in Table III require two responses each. The data for the responses to these items are numbered 1 and 2 in each table.

It will be noted that, for a few items, the proportion of successes is negative after a correction for chance had been made.

> A negative proportion may be interpreted to mean that fewer testees marked the correct answer than would have been expected by chance, probably because many of them were misinformed rather than simply uninformed with respect to the point being tested.[38]

Where the signs of the two proportions used to enter the item-analysis chart are different, there is no direct method of arriving at an accurate index of discrimination. Davis

[38] Fredrick B. Davis, Item-Analysis Data; Cambridge Massachusetts, Graduate School of Education, Harvard University, 1946, p. 3.

states that in his experience, serviceable indices can be obtained for items when the proportion of successes in the lowest twenty-seven per cent of the sample is negative by altering the proportion arbitrarily to the proportion that would result from one-half a testee having answered the item correctly.

In this study such a procedure was followed by entering the item-analysis chart at one per cent when a negative proportion occurred. The discrimination index thus obtained is always an underestimate. This fact is noted in Tables II and III by the addition of a plus superscript to the discrimination index, Flanagan's "r" and the coefficient of forecasting efficiency (E).

TABLE II

ITEM-ANALYSIS DATA FROM THE CALIFORNIA LANGUAGE TEST (ELEMENTARY)

Item No.	Highest 27%		Lowest 27%		Davis Discrim. Index	Flanagan r	E
	R	P	R	P			
A. Capitalization							
1	92	.98	87	.93	21	.35	.06
2	89	.95	65	.69	28	.43	.10
3	81	.86	72	.77	8	.13	.01
4	90	.96	56	.60	37	.54	.16
5^1	93	.99	75	.81	35	.52	.15
$_2$	88	.94	67	.72	24	.37	.07
6	73	.78	43	.46	21	.34	.06
7	93	.99	74	.80	36	.55	.16
8	90	.95	58	.64	34	.51	.14
9	90	.95	58	.64	34	.51	.14
10	92	.98	66	.73	35	.52	.15
11	84	.89	67	.71	17	.27	.04
12	71	.76	36	.41	23	.36	.07
13	48	.51	20	.23	19	.31	.05
14^1	82	.87	59	.67	17	.27	.04
$_2$	74	.79	31	.35	30	.45	.11
15	85	.90	64	.73	17	.27	.04

TABLE II (continued)

Item No.	Highest 27%		Lowest 27%		Davis Discrim. Index	Flanagan r	E
	R	P	R	P			
B. Punctuation							
16	83	.90	55	.59	24	.37	.07
17	85	.89	56	.49	30	.45	.11
18	72	.71	50	.41	19	.31	.05
19	75	.79	39	.27	52	.70	.29
20	54	.47	16	.17	22	.35	.06
21	74	.74	39	.27	31	.46	.11
22	87	.90	44	.47	33	.50	.13
23	57	.51	19	.24	18	.29	.04
24	89	.93	63	.62	29	.43	.10
25	85	.89	44	.46	33	.50	.13
26	84	.87	29	.18	52	.69	.28
27	87	.89	54	.57	26	.40	.08
28_1	61	.57	21	.08	39	.56	.17
28_2	29	.30	5	.04	30	.46	.11
29	81	.83	44	.45	27	.41	.09
30	88	.89	46	.46	33	.50	.13
C. Words and Sentences							
31	93	.98	78	.66	40	.59	.19
32	94	1.00	85	.81	35	.52	.15
33	93	.98	90	.91	17	.27	.04
34	84	.79	84	.79	0	.00	.00
35	94	1.00	91	.94	19	.31	.05
36	93	.98	78	.66	40	.59	.19
37	94	1.00	69	.47	57	.74	.33
38	70	.49	42	.11	55	.71	.30
39	93	.98	89	.89	20	.32	.05
40	93	.98	82	.74	34	.51	.14
41	81	.72	67	.43	19	.32	.05
42	90	.91	59	.26	49	.66	.25
43	88	.87	53	.13	59	.72	.31
44	85	.81	38	-.18	75$^+$	.86$^+$	.49$^+$
45	87	.85	48	.04	67	.80	.40
46	90	.91	69	.53	31	.46	.11
47	93	.98	47	.04	76	.88	.52
48	89	.89	72	.60	24	.38	.07
49	85	.81	35	-.21	75$^+$	.86$^+$	.49$^+$
50	64	.39	22	-.49	55$^+$	.72$^+$	.31$^+$

Each item that correlates positively with the final score contributes to the final measurement, to some extent, regardless of how small the correlation might be. The higher the correlation,

the more efficient is the item for general achievement test purposes.

The question is immediately raised as to how high the relationship should be between the item and the criterion or, in this study, the final score.

Items which exhibit either zero or negative correlations with the total score do not contribute to the final measurement and therefore have no value as general achievement test items.

The point at which a positively correlating item should be removed from a general achievement test cannot be stated with any finality. It depends upon the author's ability to produce another item which measures the same element yet correlates more highly with the final score. As indicated in the study by Lawshe and Mayer,[39] reliability of measurement might be raised by removing seriously defective items or by adding items which exhibit high correlations with the total score.

Davis states:

> Items with discrimination indices above 20 will ordinarily be found to have sufficient discriminating power for use in most achievement and aptitude tests.[40]

The discrimination index of 20 is equal to a product-moment r of .33 and a coefficient of forecasting efficiency in predicting the final score of 5 per cent or more, to be of

[39] Lawshe and Mayer, op. cit., p. 277

[40] Davis, op. cit., p. 15

value as general achievement test items. Presumably, those of less than 5 per cent efficiency might be replaced. This constitutes the basis upon which items of the California Language Tests were judged in this study.

It is understood that such judgments must be made only with reference to the group of pupils actually tested.

> Validity values derived for test items have specific reference only to the groups of subjects actually involved, or to groups in which the criterion ability is very similarly distributed. An item cannot be deemed to possess a certain validity per se; the value obtained may differ very widely from that obtained for another, especially if the groups differ widely from each other, either as to the average level or as to the variability of the trait concerned.[41]

The following conclusions, based on the data summarized in Table II for the California Language Test, Elementary, appear to be justified.

1. Item 34 with a discrimination index of zero, should be removed from the test. This item fails to discriminate between levels of ability of testees in the group of large urban and small urban Alberta grade four pupils.

2. Items 3, 11, 13, 14(1), 15, 18, 23, 33, 35, 39 and 41 should be replaced, if possible. These items have discrimination indices of 20 or less and are 5 per cent efficient in predicting the final score. Five per cent or less of the pupils responding correctly to these items

[41] John A. Long and Peter Sandiford, et al., The Validation of Test Items. Toronto, Dept. of Education Res., University of Toronto, 1935, p. 117

will receive higher scores than those who respond incorrectly.

3. Items 37, 38, 43, 44, 45, 47, 49 and 50 have the highest discrimination indices in the test. These items have indices of 55 or more and are more than 30 per cent efficient in predicting the final score.

4. Of a total of 53 test items in the California Language Test, Elementary, 12 or 23 per cent fail to meet the requirements for general achievement items as suggested by Davis.

TABLE III

ITEM-ANALYSIS DATA FROM THE CALIFORNIA LANGUAGE TEST (INTERMEDIATE)

Item No.	Highest 27%		Lowest 27%		Davis Discrim. Index	Flanagan r	E
	R	P	R	P			
A. Capitalization							
1	91	.99	84	.91	23	.37	.07
2	92	1.00	72	.78	37	.57	.18
3	84	.91	43	.47	35	.53	.15
4	91	.99	87	.95	16	.25	.03
5	92	1.00	77	.84	32	.48	.12
6 1	88	.96	76	.78	24	.39	.08
2	89	.97	76	.83	21	.34	.06
7	88	.96	54	.59	37	.56	.17
8	88	.96	77	.84	19	.31	.05
9	88	.96	55	.60	37	.55	.16
10	92	1.00	85	.92	22	.35	.06
11	91	.99	69	.75	40	.58	.19
12 1	92	1.00	73	.79	36	.54	.16
2	92	1.00	76	.83	33	.50	.13
13 1	83	.90	85	.92	-3	-.05	.00
2	89	.97	67	.82	24	.39	.08
14	73	.79	54	.59	14	.23	.03
15	67	.82	36	.39	29	.45	.11

TABLE III (continued)

Item No.	Highest 27%		Lowest 27%		Davis Discrim. Index	Flanagan r	E
	R	P	R	P			
B. Punctuation							
16	71	.72	70	.74	-1	-.02	.00
17	91	.99	73	.79	36	.54	.16
18	83	.90	58	.58	26	.41	.09
19	70	.70	45	.40	19	.31	.05
20 1	72	.73	46	.41	20	.32	.05
2	66	.65	28	.16	34	.51	.14
21	84	.90	62	.64	23	.37	.07
22	87	.94	62	.64	29	.43	.10
23	80	.79	47	.42	25	.39	.08
24 1	75	.77	43	.38	27	.41	.09
2	60	.57	22	.07	42	.60	.20
25	87	.93	65	.72	22	.35	.06
26	86	.92	59	.63	26	.41	.09
27	89	.96	61	.66	32	.48	.12
28	92	1.00	73	.86	30	.45	.11
29	49	.42	24	.12	24	.38	.07
30	90	.97	70	.86	19	.31	.05
31	82	.89	59	.70	17	.27	.04
32	64	.64	56	.64	0	.00	.00
33	35	.26	11	-.06	41$^+$	.59$^+$	.19$^+$
34	74	.85	52	.69	15	.23	.03
35	81	.95	57	.77	21	.34	.06
C. Words and Sentences							
36	92	1.00	83	.80	36	.54	.16
37	86	.87	77	.67	17	.27	.04
38	90	.96	81	.76	25	.40	.08
39	89	.93	71	.54	34	.51	.14
40	85	.85	57	.24	43	.61	.21
41	91	.98	86	.87	22	.35	.06
42	86	.87	72	.57	23	.37	.07
43	69	.50	46	.00	55	.73	.31
44	76	.65	57	.24	27	.41	.09
45	92	1.00	86	.87	29	.45	.11
46	91	.98	69	.50	50	.68	.27
47	92	1.00	73	.59	50	.68	.27
48	92	1.00	84	.90	25	.40	.08
49	91	.98	75	.69	37	.55	.16
50	92	1.00	77	.73	42	.60	.20
51	92	1.00	77	.75	40	.58	.19
52	91	.98	75	.70	37	.55	.16
53	86	.87	51	.16	51	.69	.28
54	88	.91	67	.52	32	.48	.12
55	78	.70	33	-.25	68$^+$	.81$^+$	.41$^+$

TABLE III (continued)

Item No.	Highest 27%		Lowest 27%		Davis Discrim. Index	Flanagan r	E
	R	P	R	P			
D. Parts of	Speech						
56	87	.93	44	.35	46	.64	.23
57	90	.97	55	.50	46	.64	.23
58	81	.85	25	.09	57	.74	.33
59	43	.33	16	-.03	46^{+}	.64^{+}	.23^{+}
60	87	.93	60	.57	32	.48	.12
61	34	.21	20	.02	30	.46	.11
62	72	.73	12	-.09	69^{+}	.81^{+}	.41^{+}
63	88	.95	34	.22	58	.75	.34
64	82	.86	39	.29	40	.58	.19
65	85	.90	36	.25	48	.66	.25
66	57	.52	22	.05	41	.59	.19
67	88	.95	28	.13	66	.80	.40
68	89	.96	61	.59	37	.55	.16
69	89	.96	41	.32	54	.72	.30
70	92	1.00	59	.58	51	.69	.28
71	8	-.14	7	-.15	0	.00	.00
72	81	.86	29	.16	50	.68	.27
73	79	.84	26	.12	52	.70	.29
74	45	.37	20	.03	38	.56	.17
75	89	.97	54	.49	47	.64	.23

The following conclusions are based on the data presented in Table III and have reference to Alberta grade seven pupils of large and small urban areas.

1. Items 13(1) and 16 discriminate negatively and therefore are of no value as general achievement test items.
2. Items 32 and 71, with discrimination indices of zero, fail to discriminate between grade seven Alberta pupils of high and low levels of ability and should be removed from the test.
3. Items 4, 8, 14, 19, 20 (1), 30, 31, 34 and 37 have discrimination indices of 20 or less and are 5 per cent or

less efficient in predicting the criterion. According to the suggestion of Davis, these should be replaced by items which have higher discriminating power.

4. Items 43, 58, 62, 63, and 67 have the highest discrimination indices in the test. Their indices of 55 or more make them more than 30 per cent efficient in predicting total score.

5. Of the 80 items in the intermediate test, 13 or 16 per cent fail to meet the requirements for achievement test items as proposed by F. B. Davis.

Table IV contains the percentage in each sub-test of the elementary and intermediate tests which, on the basis of the test results of Alberta pupils, may be considered inefficient to the extent that they should be replaced as achievement test items.

TABLE IV

PERCENTAGE OF INEFFICIENT ACHIEVEMENT TEST ITEMS IN SUB-TESTS OF THE CALIFORNIA LANGUAGE TESTS (ELEMENTARY AND INTERMEDIATE)

Test 5 Sub-tests	Elementary		Intermediate	
	No.	%	No.	%
A. Capitalization	5	29	4	22
B. Punctuation	2	13	7	32
C. Words and Sentences	5	25	1	5
D. Parts of Speech	-	-	1	5

3. Summary of Findings

1. Interpretation of responses in Section A of both tests would be facilitated by providing a place where a testee might indicate, positively, that no capitalization was needed.

2. A testee should be penalized for over-capitalization and over-punctuation.

3. Twelve of 53 test items or 23 per cent of the items in the California Language Test, Elementary, should be replaced as achievement test items. As previously stated, items such as these, with discrimination indices of 20 or less, have insufficient discriminating power for use in achievement tests.

4. In the elementary test, Section B - Punctuation, contains the fewest inefficient items (13 per cent), while Section A - Capitalization, contains the most (29 per cent).

5. Thirteen or 16 per cent of the intermediate test items should be replaced as achievement test items because of their lack of discriminating power.

6. In the intermediate test, Section C - Words and Sentences, and Section D - Parts of Speech, contain the smallest percentage of inefficient items (5 per cent in each), while Section B - Punctuation, contains the largest percentage (32 per cent).

CHAPTER VI

DATA ON ITEM DIFFICULTY

1. Distributions of Item Difficulty

If the only discrimination indices used in selecting items for a test are based on the total score, it is desirable that items having the highest indices be chosen. But if maximum efficiency of measurement is to be attained the requirements of distribution and pattern of item difficulty must also be taken into account.

Items, measuring ability at one level with high discriminating power, may have very low discriminating power at a different level. The fact is well known that the discriminating power of an item is greatest when it is used in a sample in which fifty per cent of the testees respond to the item correctly.

In construction of achievement tests, where maximum discrimination among all testees is desired, the optimum shape of the distribution becomes a function of the intercorrelations of the items. Flanagan, in 1939 stated:

> To obtain maximum discrimination among the individuals in a particular group, a test should be composed of items all of which are of fifty per cent difficulty provided that the intercorrelations of the items are zero ... It can be shown that a rectangular distribution of item difficulty ... is necessary to obtain maximum discrimination among members of the group provided that the intercorrelations between all of the items are unity.[42]

[42] Flanagan, op. cit., pp. 675-676

To illustrate Flanagan's line of reasoning, we will use as an example, a test of ten items with a sample of one hundred testees. The maximum number of discriminations that may be made by an item for one hundred testees is 50 times 50 or 2,500. This occurs when the item is at a fifty per cent difficulty level. When an item is at another level of difficulty the number of discriminations is reduced. For example, an item of thirty per cent difficulty will make only 2,100 discriminations. If the ten items are uncorrelated then they are able to make a maximum of 25,000 discriminations when each item is at a fifty per cent level of difficulty.

Since the maximum number of discriminations that can be made by a single item with a sample of one hundred testees is 2,500, the maximum number of discriminations that may be made by ten perfectly correlated items of the same level of difficulty is also 2,500. Under these conditions, the other nine items provide no additional measurement. If the "ten items are evenly spaced from a difficulty level of five per cent to one of ninety-five per cent, a maximum of 4,525 discriminations may be made."[43]

Since intercorrelations of unity or zero among test items would seldom if ever be found, the two hypothetical conditions stated by Flanagan may be considered to be limiting cases. The conditions encountered in practice are usually between these

[43] Davis, op. cit., p 25

two limits.

Where item intercorrelations are low and positive, maximum discrimination will be obtained when the items cluster about the fifty per cent level of difficulty. However, as the "intercorrelations of test items are increased, the distribution of item difficulty should become more platykurtic with a smaller proportion clustering near the 50% difficulty level."[44]

In practice, the intercorrelations of a set of test items are seldom known. However, some indication of the relationship among items is provided by the size of the average discrimination index.

> If the discrimination indices of a set of well-edited items tend to run low, the reason is usually that the items are measuring rather specific elements and are not highly intercorrelated.[45]

In such cases, the most effective test would contain items covering a wide range of difficulty with the majority near the fifty per cent level. Where the discrimination indices are high, it might be assumed that the item intercorrelations are likewise high. In these circumstances, the most effective test would consist of items whose distribution of difficulty is nearly rectangular.

The mean discrimination indices of the California Language Tests, Elementary and Intermediate, and their component

[44] _Ibid_.

[45] _Ibid_. p.26

sub-tests are summarized in Table V. These averages were computed by simply adding, algebraically, the Davis discrimination indices and by dividing the sum by the number of test items.

TABLE V

MEAN DISCRIMINATION INDICES FOR THE CALIFORNIA LANGUAGE TESTS, ELEMENTARY AND INTERMEDIATE

Elementary		Intermediate	
	Mean Discrim. Index		Mean Discrim. Index
All Items	33	All Items	36
Sub-tests:		Sub-tests:	
A. Capitalization	26	A. Capitalization	32
B. Punctuation	31	B. Punctuation	24
C. Words and Sentences	42	C. Words and Sentences	37
		D. Parts of Speech	52

The fact that the average discrimination index for all items is low in both tests may be due to the fact that the items are measuring rather specific elements that are not highly intercorrelated. If this is the case, it might be expected that the items would be distributed over a wide range of difficulty with the majority near the fifty per cent level of difficulty. Sub-tests likewise have low average discrimination indices and should have distributions of item difficulty similar to those of the total test. Section C. of the elementary test and Section D. of the intermediate, having somewhat higher average discrimination indices than other sub-

tests, might be expected to have more platykurtic distributions of item difficulty.

In this study, the percentage of success is used as an index of item difficulty. It is the percentage of pupils in the sample who responded correctly to a test item. The percentages of pupils responding successfully, unsuccessfully, and making no response are summarized in Tables VI and VII. The distributions of item difficulty are illustrated by histograms in Figures 1 to 9.

TABLE VI

PERCENTAGE OF SUCCESSES AND FAILURES ON ITEMS OF THE CALIFORNIA LANGUAGE TEST, ELEMENTARY

Item No.	% Successful	% Unsuccessful	% No. Response
A. Capitalization			
1	94	6	0
2	86	14	0
3	69	31	0
4	88	12	0
5^{1}	89	11	0
$_{2}$	86	14	0
6	74	26	0
7	93	7	0
8	74	26	0
9	78	22	0
10	88	12	0
11	87	13	0
12	63	34	3
13	45	51	4
14^{1}	72	24	4
$_{2}$	50	46	4
15	77	19	4
Mean % Success = 77			

TABLE VI (continued)

Item No.	% Successful	% Unsuccessful	% No. Response
B. Punctuation			
16	68	32	0
17	71	28	1
18	53	43	4
19	60	39	1
20	68	31	1
21	55	45	0
22	70	28	2
23	66	33	1
24	76	24	0
25	72	26	2
26	54	40	6
27	64	24	12
28 1	36	33	31
2	12	59	29
29	63	27	10
30	59	28	13
Mean % Success = 56			
C. Words and Sentences			
31	90	10	0
32	98	2	0
33	99	1	0
34	89	11	0
35	98	2	0
36	89	11	0
37	84	16	0
38	64	36	0
39	99	1	0
40	97	3	0
41	84	16	0
42	87	13	0
43	79	21	0
44	61	39	0
45	82	18	0
46	95	5	0
47	88	11	1
48	91	8	1
49	66	33	1
50	43	53	4
Mean % of Success = 84			
Mean Percentage of Success = 74 (all items)			

Figure 1 illustrates the distribution of difficulty in terms of percentage of success, for items of the California Language Test, Elementary. The data for this distribution are recorded in Table VI.

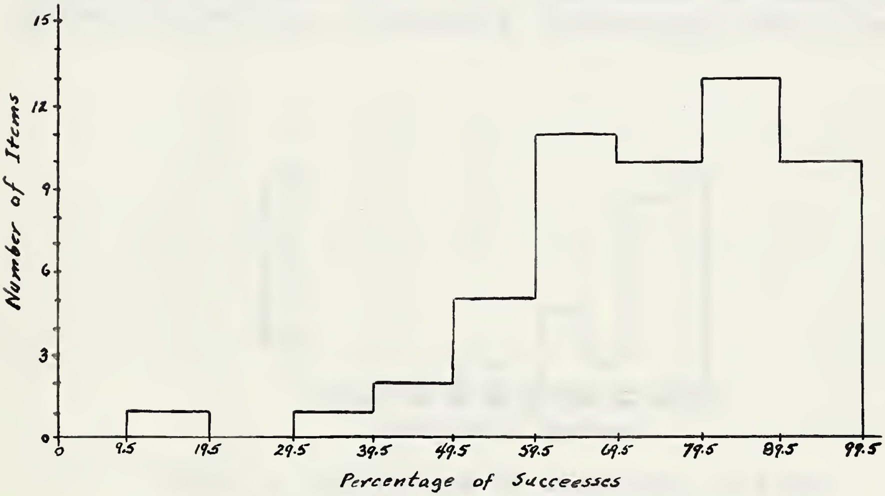

Figure 1. Distribution of Percentage of Successes on Items of the California Language Test, Elementary

The distribution of percentages of success, as illustrated in Figure 1, indicates that the range of difficulty of the items is satisfactory but that the test contains too many relatively easy items. With a mean percentage of success of 74, it contains too many items at the lower levels of difficulty and too few near the fifty per cent difficulty level.

Figures 2, 3 and 4 illustrate the distributions of item difficulty for sub-tests of the California Language Test, Elementary.

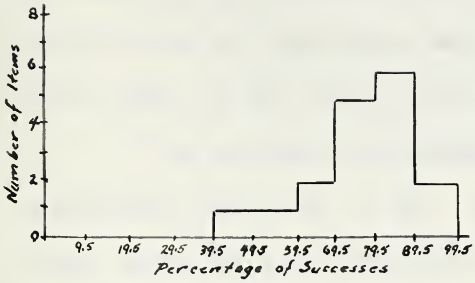

Figure 2. Percentage of Successes on Items of the Capitalization Test, Elementary

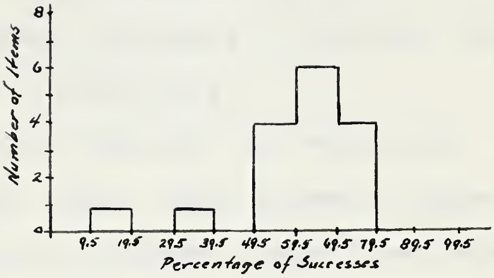

Figure 3. Percentage of Successes on Items of the Punctuation Test, Elementary

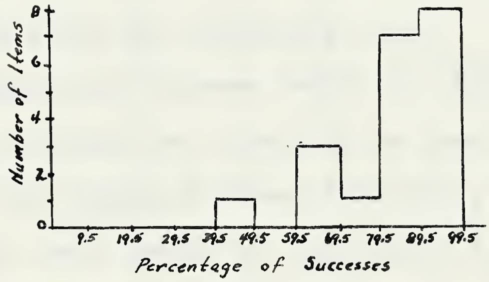

Figure 4. Percentage of Successes on Items of the Words and Sentences Test, Elementary

The distributions of item difficulty of sub-tests, as illustrated in Figures 2, 3 and 4, indicate that the range of item difficulty in each case is too low for adequate measurement of general achievement of the sample used in this study.

The capitalization test, with a mean percentage of success of 77, contains too many items at the lower levels of difficulty, too few at the upper levels and too few near the fifty per cent difficulty level.

The punctuation test, with a mean percentage of success of 56, provides the best distribution of item difficulty of the three sub-tests. However, it contains too few items at the lower levels of difficulty.

The average discrimination index for the words and sentences sub-test is 42. This index, being somewhat higher than those of the other sub-tests, suggests higher item intercorrelations in the sub-test. As stated earlier, an increase in intercorrelation of items should be accompanied by more platykurtic distributions of item difficulty. But this is not at all apparent in the distribution of percentage of success for the words and sentences test.

Discriminations between levels of ability in the words and sentences sub-test are likely to be less than optimal because of the low range of item difficulty, the concentration of items at the lower levels of difficulty and the peaked distribution of item difficulty.

Table VII summarizes the percentages of pupils responding successfully, unsuccessfully and making no response to the eighty items of the California Language Test, Intermediate.

TABLE VII

PERCENTAGE OF SUCCESSES AND FAILURES ON ITEMS OF THE CALIFORNIA LANGUAGE TEST, INTERMEDIATE

Item No.	% Successful	% Unsuccessful	% No. Response
A. Capitalization			
1	96	4	0
2	90	10	0
3	69	31	0
4	97	3	0
5	94	6	0
6^1	91	9	0
$_2$	91	9	0
7	79	21	0
8	92	7	1
9	80	20	0
10	98	1	1
11	89	10	1
12^1	93	6	1
$_2$	93	6	1
13^1	85	14	1
$_2$	89	10	1
14	71	27	2
15	55	42	3
Mean % Success = 86			
B. Punctuation			
16	74	26	0
17	94	6	0
18	85	15	0
19	64	35	1
20^1	72	28	0
$_2$	48	50	2
21	81	18	1
22	79	19	2
23	72	26	2
24^1	69	29	2
$_2$	46	52	2
25	86	11	3
26	82	15	3
27	86	11	3
28	91	5	4
29	37	59	4
30	90	5	5
31	81	12	7
32	60	31	9
33	20	69	11
34	69	16	15
35	79	6	15
Mean % Success = 78			

TABLE VII (continued)

Item No.	% Successful	% Unsuccessful	% No. Response
C. Words and Sentences			
36	93	7	0
37	88	12	0
38	94	6	0
39	88	12	0
40	76	24	0
41	97	3	0
42	84	16	0
43	60	40	0
44	72	28	0
45	98	2	0
46	89	11	0
47	91	9	0
48	97	2	1
49	92	7	1
50	93	6	1
51	95	4	1
52	93	6	1
53	73	26	1
54	89	10	1
55	59	40	1
Mean % Success = 86			
D. Parts of Speech			
56	73	27	0
57	86	14	0
58	60	40	0
59	34	66	0
60	85	15	0
61	29	71	0
62	43	57	0
63	70	30	0
64	69	30	1
65	73	26	1
66	41	59	0
67	64	36	0
68	86	14	0
69	77	22	1
70	85	14	1
71	10	89	1
72	67	32	1
73	60	39	1
74	31	68	1
75	82	17	1
Mean % Success = 61			
Mean Percentage of Success = 77 (all items)			

Figure 5 illustrates the distribution of difficulty for items of the California Language Test, Intermediate. The data for this distribution are found in Table VII.

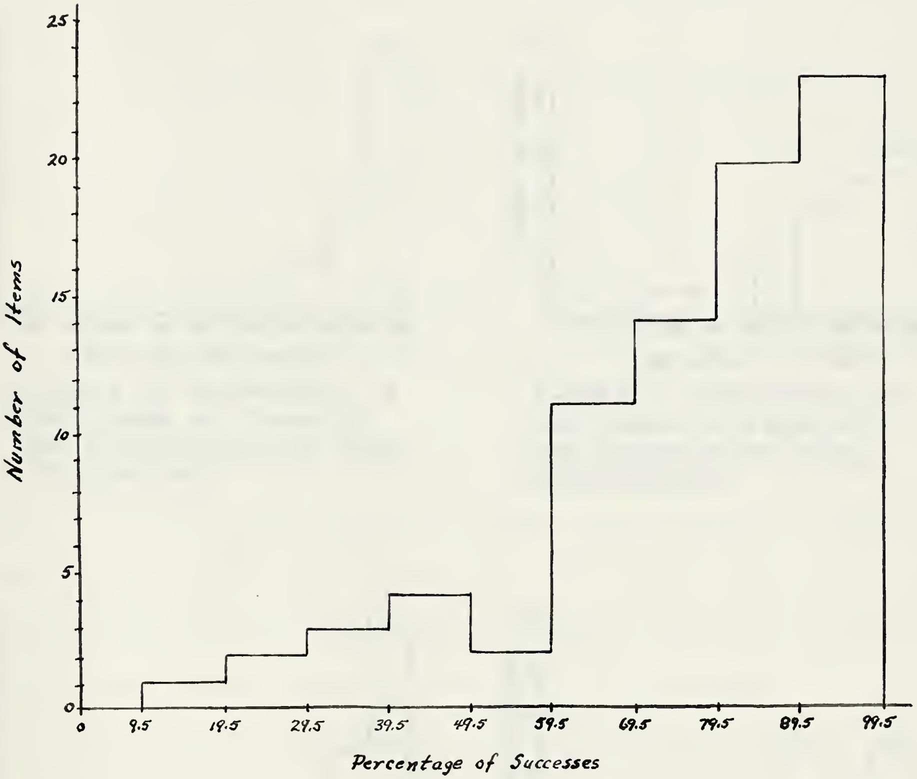

Figure 5. Distribution of Percentage of Successes on Items of the California Language Test, Intermediate

The distribution of item difficulty in the California Language Test, Intermediate, as presented in Figure 5, indicates that there is an adequate range of item difficulty but that the test, like the elementary test, contains too many relatively easy items. The mean percentage of success for the test is 77.

The distributions of item difficulty for the capitalization, punctuation, words and sentences, and parts of speech sub-tests are presented in Figures 6, 7, 8 and 9.

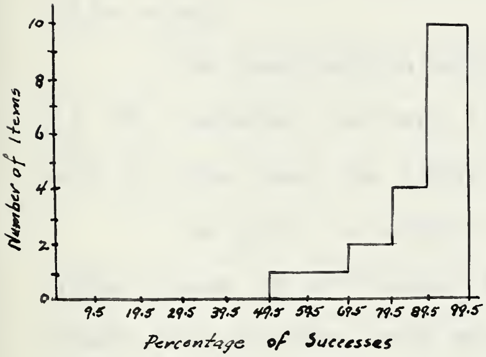

Figure 6, Percentage of Successes on Items of the Capitalization Test, Intermediate

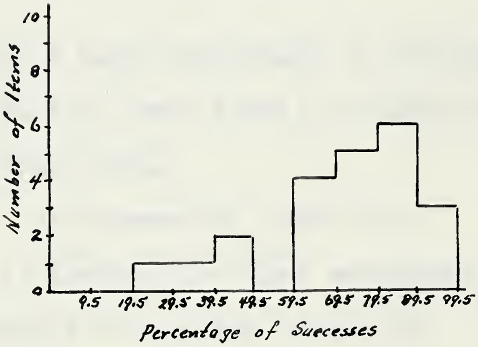

Figure 7. Percentage of Successes on Items of the Punctuation Test, Intermediate

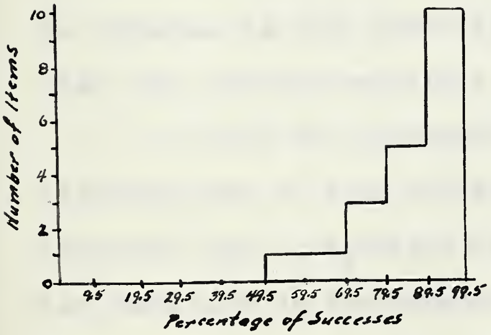

Figure 8. Percentage of Successes on Items of the Words and Sentences Test, Intermediate

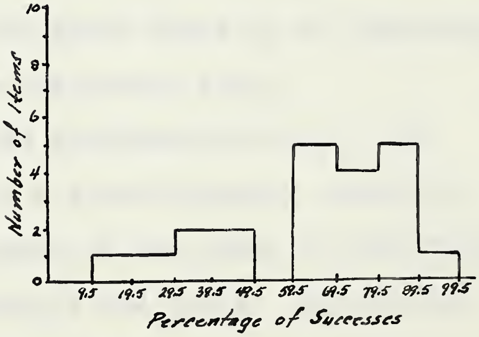

Figure 9. Percentage of Successes on Items of the Parts of Speech Test, Intermediate

The capitalization and words and sentences sub-tests, each with a mean percentage of success of 86, have very similar distributions of item difficulty. Both sub-tests contain too many relatively easy items and have insufficient range of item difficulty for the sample of pupils used in this study.

The punctuation test, with a mean percentage of success of 78, contains too many items at the lower levels of difficulty and too few near the fifty per cent level.

The parts of speech test is composed of items of a wider range of difficulty than is present in other sub-tests. There are, however, too many items at the lower levels of difficulty.

With the highest average discrimination index of the four sub-tests, the parts of speech test appears to have the most platykurtic distribution of item difficulty. This should be expected in such cases as this where there is an indication that item intercorrelations are relatively high.

In both the elementary and intermediate tests, the distributions of item difficulty are sufficiently peaked to indicate that a selection of items, on the basis of difficulty, had been made in the construction of the tests. But because of the large number of easy items in each test, the distributions are negatively skewed. This will have the effect of reducing the discriminating power of the tests at the upper levels of ability in the sample.

Both the elementary and intermediate tests were used on samples of the lowest grades for which they were designed. The elementary test, intended for grades four, five and six, was administered to a grade four sample, while the intermediate test intended for grades seven, eight and nine, was administered to a grade seven group. Since the tests contain an unduly large number of relatively easy items for grade four and seven pupils, it might be expected that the distributions of item difficulty would be still more seriously skewed if the tests were employed at their upper grade levels. This will result in less effective measurement of general achievement when the tests are administered to grades six and nine pupils.

2. Pattern of Item Difficulty

The current practice in test construction is to present items covering a wide range of difficulty, arranged in ascending order of difficulty from the very easy to the most difficult . This makes it possible for the less able pupil to respond to items within his level of ability without being discouraged early in the test by items of prohibitive difficulty. The most difficult items, at the end of the test, differentiate between pupils of superior ability.

Where an achievement test contains a number of sub-tests, an arrangement of items in ascending order of difficulty might be expected in each sub-test. They should begin with their easiest items and end with the most difficult.

A control of levels of item difficulty provides effective measurement of achievement over a wide range of ability. It may also contribute to the motivation of pupils taking the test by arousing confidence in the less able and by presenting challenging situations to superior pupils.

The patterns of item difficulty for the elementary test, are presented graphically in Figures 10, 11 and 12. The datum for each item, as recorded in Table VI, is presented in the order that the item appears on the test.

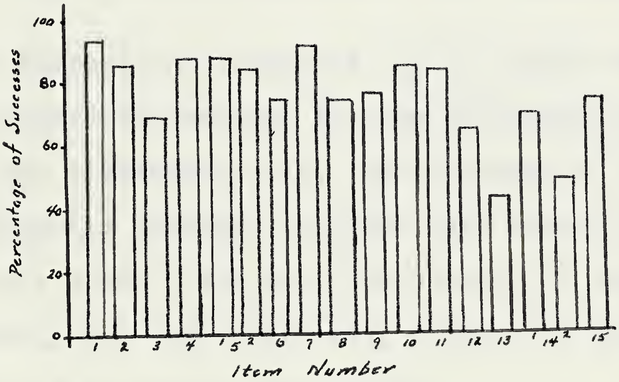

Figure 10. Percentage of Successes on Items of the Capitalization Test, Elementary

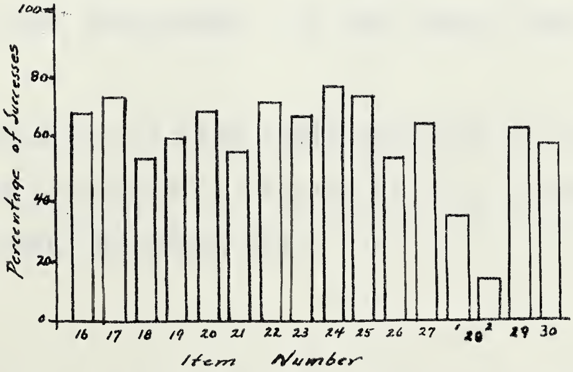

Figure 11. Percentage of Successes on Items of the Punctuation Test, Elementary

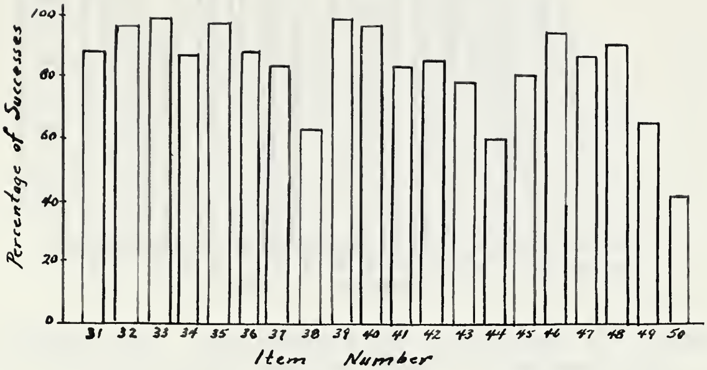

Figure 12. Percentage of Successes on Items of the Words and Sentences Test, Elementary

An inspection of Figures 10, 11 and 12 reveals a tendency toward an increase in item difficulty in the sub-tests of the elementary test. This pattern of item difficulty is only slightly apparent from the test results of the sample used in this study. The large proportion of items of approximately the same difficulty causes the pattern of item difficulty to be barely descernible.

Items in the sub-tests may have been arranged in ascending order of difficulty for the normalizing sample but evidence of this arrangement for the sample used in this study is minimal.

The patterns of item difficulty for the intermediate sub-tests are presented in Figures 13, 14, 15 and 16. The data are recorded in Table VII.

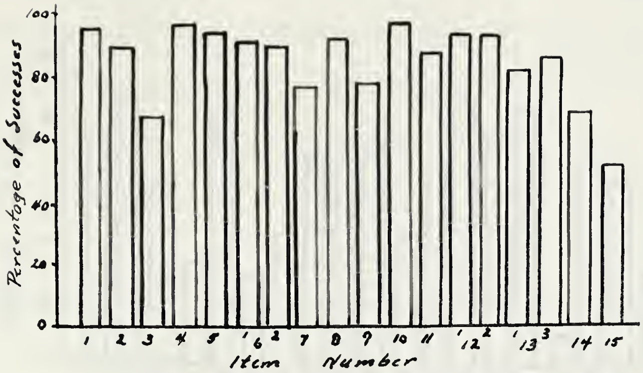

Figure 13. Percentage of Successes on Items of the Capitalization Test, Intermediate

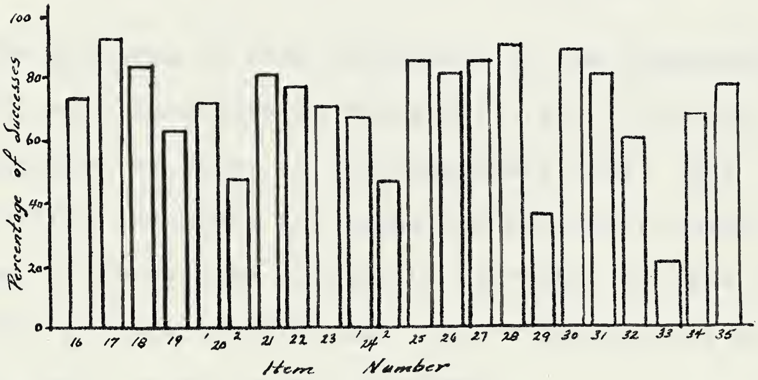

Figure 14. Percentage of Successes on Items of the Punctuation Test, Intermediate

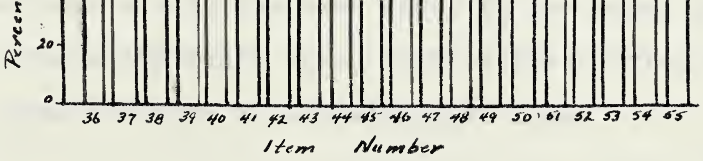

Figure 15. Percentage of Successes on Items of the Words and Sentences Test, Intermediate

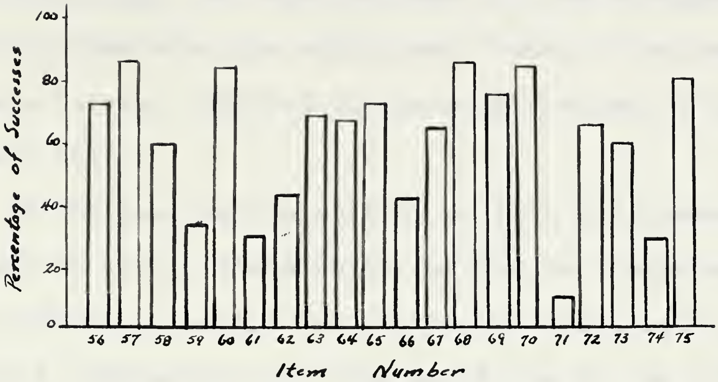

Figure 16. Percentage of Successes on Items of the Parts of Speech Test, Intermediate

The patterns of item difficulty of the intermediate sub-tests, as illustrated in Figures 13, 14, 15 and 16, are not so apparent as those of the elementary test. The capitalization and words and sentences sub-tests appear to have items of increasing difficulty only near the end of the tests. The pattern of difficulty in the punctuation and parts of speech tests is not at all apparent.

The irregularities in levels of difficulty in the parts of speech sub-test may be due, in part, to the fact that pupils are asked to identify parts of speech in a single complete sentence. The construction of the sentence is such that a relative pronoun is placed immediately before a verb. And as the items are in the same order as the words in the sentence, a more difficult item, such as identifying a relative pronoun, appears before a less difficult item.

It appears doubtful that adequate control of levels of difficulty and the usual pattern of item arrangement may be accomplished when the additional factor of sentence structure is also involved in the construction of the parts of speech test.

In the punctuation section of both the elementary and intermediate tests, items which require two responses have higher difficulty levels for the second response in each case. A possible explanation for this might lie in the fact that the directions to the student are not sufficiently explicit for them to expect items requiring two responses. The only indication that more than one response may be required is given in the last sentence of the directions which reads, "Use the same answer row to show all punctuation needed at any one number in the story."

This may have the effect of making items requiring double responses, more difficult than they would be if the elements being tested, were presented separately.

3. Summary of Findings

1. The distributions of item difficulty for both the elementary and intermediate tests are peaked, indicating that a selection of items on the basis of difficulty, has been carried out in the construction of the tests.
2. The distributions of item difficulty for both tests are not well adjusted to the abilities of Alberta grade four and seven pupils of large and small urban centres.

Because of the large proportion of relatively easy items, all distributions of difficulty are negatively skewed. This is most pronounced in the capitalization and words and sentences sub-tests of the intermediate test, and in the words and sentences section of the elementary test.

3. There is a slight tendency for the items in the sub-tests of the elementary test to be arranged in ascending order of difficulty, from the least to the most difficult.

4. A similar pattern of item difficulty in the intermediate test is somewhat apparent in the capitalization and words and sentences sub-tests. In the punctuation and parts of speech tests, items are arranged in an irregular fashion with respect to item difficulty.

The irregularities in the punctuation section may be due to difficult items resulting from the lack of explicit directions to pupils. These directions may not enable pupils to expect to encounter items requiring double responses. The irregularities in levels of difficulty in the parts of speech test is undoubtably due to the fact that pupils are required to recognize parts of speech in a single complete sentence. As a relative pronoun precedes a verb in the sentence, a difficult item, at times, precedes a less difficult item.

CHAPTER VII

SUMMARY OF FINDINGS AND CONCLUSIONS

1. Findings

The purpose of this study was to determine the effectiveness of the California Language Tests for samples of Alberta grades four and seven pupils, by the use of item-analysis techniques. An examination of the data presented in this study appears to justify the following:

1. In the capitalization section of both the elementary and intermediate tests, an ambiguity arises when no response is made to an item. This might be interpreted to mean that the testee, for some reason, failed to respond to the item or that he decided that no capitalization was necessary. The interpretation of responses in this section would be facilitated by providing a place where a testee might indicate, positively, that no capitalization is required.
2. The effectiveness of the items in the capitalization and punctuation sub-tests of the elementary and intermediate tests, is reduced when equal credit is given to pupils marking only the correct response, and to others marking several, of which only one is correct. These tests might be improved by penalizing pupils for over-capitalization and over-punctuation.

3. One item in the elementary test and four in the intermediate test, with either zero or negative discrimination indices, fail to discriminate between levels of ability of the sample of pupils used in this study. These items do not contribute to the measurement of language achievement.

4. Twenty-three per cent of the items of the elementary test and sixteen per cent of the intermediate test items have insufficient discriminating power to meet the requirements for achievement test items as proposed by F. B. Davis.

5. In the elementary test, the fewest inefficient items are found in the punctuation test. The capitalization test contains the most. The words and sentences, and parts of speech sections of the intermediate test have the fewest inefficient items, while the punctuation section has the most.

6. The distributions of item difficulty for both the elementary and intermediate tests are sufficiently peaked to indicate that a selection of items on the basis of difficulty, had been carried out in the construction of the tests.

7. Because of the large proportion of relatively easy items, the distributions of difficulty for all tests are negatively skewed. This is most pronounced in the words and sentences sub-test of the elementary test and in the

capitalization and words and sentences sections of the intermediate test.

8. The sub-test items of the elementary test are too similar in level of difficulty to permit a clearly defined arrangement in order of difficulty. However, to the degree that there was variation in difficulty of items, this pattern has been maintained. In the intermediate test, patterns of item difficulty are even less apparent in the capitalization, and words and sentences sub-tests.

9. In the punctuation and parts of speech tests, items are arranged in an irregular fashion with respect to level of difficulty. In the punctuation test this may be due to lack of explicit directions concerning items which require two or more responses, with the consequence that such items become more difficult than they would be were the elements measured separately. Irregular levels of difficulty in the parts of speech test may be due to the fact that the testers are required to identify parts of speech in the order of occurrence in a sentence. Almost inevitably order of occurrence does not coincide with order of difficulty.

2. Conclusions

The effectiveness of the California Language Tests, Elementary and Intermediate, appears to be limited by a number of factors. These factors may be of little consequence when

considered independently but together they tend to reduce the efficiency of measurement.

1. Defects in the scoring procedure and in the method of recording answers in the capitalization and punctuation sections, reduce the effectiveness of the items. Equal credit is given to pupils, some of whom mark only the correct responses while others mark several responses to each item. There would appear to be a difference in the ability of these pupils but this is not always indicated by the final score.
2. In both tests, the Davis discrimination indices are relatively low indicating a lack of test homogeneity. As a result, some ambiguity in the interpretation of the total score may be expected.

The fact that items appear to be measuring specific elements of language skill which are not highly correlated may account for the lack of internal consistency in the test as a whole. Because of their restricted content, the sub-tests may possess greater test homogeneity and yield relatively unambiguous scores. But before one places any credence in their results he should be assured of their reliability. The test authors of the California Language Tests have not provided this information.

With the California Language Tests, the test user is provided with two sets of scores; the total score, which

may have little meaning because of the lack of internal consistency of the tests and the sub-scores, which may be the product of relatively homogeneous measures, but for which no reliability data are reported.

The current practice of attempting to produce an instrument to serve both as a diagnostic and an achievement test, reduces the possibility of effective measurement for either purpose.

3. Perhaps the greatest factor limiting the effectiveness of the California Language Tests is the unduly large proportion of relatively easy items. The distributions of item difficulty are not well adjusted to the abilities of Alberta grade four and grade seven pupils in large and small urban centres. The large proportion of relatively easy items in both tests, has the effect of reducing the power of the test to differentiate between pupils at the upper levels of ability.

4. Since the tests contain too many relatively easy items for the sample of grade four and seven pupils, it might be expected that the distributions of item difficulty would be still more seriously skewed if the tests were administered to grades six and nine pupils.

5. The California Language Tests are limited in their usefulness in that they measure but a few of the basic objectives of language instruction. Their content is restricted to the use of capital letters and punctuation

marks, and the recognition of complete sentences and parts of speech. No attempt is made to measure some of the broader outcomes such as the ability to organize and present material logically, to sustain an idea, to establish continuity of thought, and to discriminate in the selection of words. For this reason they must be regarded as tests of language usage rather than comprehensive tests of language ability. This is a weakness which is common to the language portion of most of the current achievement tests.

BIBLIOGRAPHY

Anastasi, Anne, Psychological Testing. New York, The Mcmillan Company, 1954.

Anderson, John E. "The Effect of Item Analysis upon the Discriminative Power of an Examination". Journal of Applied Psychology, Volume 19, 1935, pp. 237-244.

Brueckner, L.J. and Melby, E.O. Diagnostic and Remedial Teaching. Cambridge, Mass., The Riverside Press, 1931.

Caldwell, O.W. and Courtis, S.A. Then and Now in Education, 1845-1923. Yonkers-on-Hudson, N.Y., World Book Company, 1925.

Clark, Willis W. California Achievement Tests. A Technical Report Supplementing Information Presented in the Manuals of Directions, Los Angeles, California Test Bureau, 1951.

Conquest, George R. A Survey of Language Achievement in Grades Four and Seven in Selected Alberta Schools. Unpublished Master's Thesis, The University of Alberta, August, 1954.

Coutts, H.T. and Baker, H.S. "A Study of the Written Composition of a Representative Sample of Alberta Grade Four and Grade Seven Pupils", Alberta Journal of Educational Research, Vol. 1, No. 2, June, 1955, pp. 5-18.

Cronbach, L.J. Essentials of Psychological Testing. New York, Harper Brothers, 1949.

Davis, Frederick B. Item-Analysis Data. Cambridge, Mass., Graduate School of Education, Harvard University, c1946.

DeBoer, John J. Kaufers, Walter V. and Miller, Helen R. Teaching Secondary English. New York, McGraw-Hill Book Company, 1951.

Flanagan, John C. "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from Data at the Tails of the Distribution". Journal of Educational Psychology, Volume XXX, 1939, pp. 674-680.

Garrett, Henry E. Statistics in Psychology and Education. New York, Longmans, Green and Company, 1947.

Greene, Harry A. "English - Language, Grammar, and Composition". Encyclopedia of Educational Research, New York, The Mcmillan Company, 1952, pp. 383-395.

Greene, Harry A., Jorgensen, A.N. and Gerbich, J.R. *Measurement and Evaluation in the Elementary School*. Toronto, Longmans, Green and Company, 1947.

Grossnickle, Foster E. "Review of the Arithmetic Essentials Test". *The Fourth Mental Measurements Yearbook,* Highland Park, N.J., Gryphon Press, 1953.

Kelley, T.L. "The Selection of Upper and Lower Groups for the Validation of Test Items". *Journal of Educational Psychology*, Volume XXX, 1939, pp. 17-24.

Lannholm, Gerald V. "Review of California Language Tests", *The Fourth Mental Measurements Yearbook*, edited by O.K. Buros, Highland Park, N.J., Gryphon Press, 1953, pp. 148-149.

Lawshe, C.H. and Mayer, James S. "The Effect of Two Methods of Item Validation on Test Reliability", *Journal of Applied Psychology*, Volume 31, 1947, pp. 271-277.

Long, John A. and Sandiford, Peter. *et al*., *The Validation of Test Items*. Toronto, Department of Educational Research, University of Toronto, 1935.

Pooley, Robert C. "Review of California Language Tests", *The Fourth Mental Measurements Yearbook*, edited by O.K. Buros, Highland Park, N.J., Gryphon Press, 1953, pp. 151-152.

Reid, T.J. *A Survey of the Language Achievement of Alberta School Children in Relation to Bilingualism, Sex and Intelligence*, Unpublished Master's Thesis, the University of Alberta, September, 1954.

Tiegs, Ernest W. and Clark, Willis W. *California Language Test Manual*, Los Angeles, California Test Bureau, 1950.

Thorndike, R.L. and Hagen, E. *Measurement and Evaluation in Psychology and Education*, New York, John Wiley and Sons, Inc., 1955.